

UNITED STATES PATENT APPLICATION

FOR

STACKABLE ELECTRONIC ASSEMBLY

Inventors
Mark Ellsberry
Charles E. Schmitz
Chi She Chen

Prepared by:

Snell & Wilmer, L.L.P.
1920 Main St., Suite 1200
Irvine, California 92614-7230
(949) 253-2700

STACKABLE ELECTRONIC ASSEMBLY

FIELD

[0001] Various embodiments of the invention pertain to stackable electronic assemblies. At least one embodiment of the invention pertains to a memory module using these stackable assemblies that enables higher memory densities.

DESCRIPTION OF RELATED ART

[0002] Semiconductor dice, such as memory dice, are often packaged for protection and ease of use. One type of package is a ball grid array package (BGA) in which a semiconductor die is mounted on a substrate. The semiconductor die has a plurality of bond pads that are electrically connected, via wires bonded between the metal traces on the substrate and on the die. The traces on the substrate terminate at contact pads where conductive elements, such as solder balls, are attached. The BGA package can then be mounted on a circuit board and electrically connected via the conductive elements. The BGA package may be electrically connected to metal traces on the circuit board in various ways, including applying heat or ultrasound to the conductive elements (e.g., solder balls).

[0003] One consideration when packaging semiconductor dice is thermal cracking. Thermal cracking may occur when the differences between the thermal coefficients of expansion for two materials causes solder points between them to crack. For instance, if a semiconductor die was directly soldered onto a substrate having a substantially different coefficient of expansion, temperature cycling may eventually cause soldered points, such as solder balls, to crack thus causing an electrical discontinuities. To solve this problem when packaging semiconductor dice, the area between the semiconductor die and the substrate is often underfilled, with epoxies and/or other materials, to assist in preventing thermal cracking.

[0004] Additionally, semiconductor dice are typically quite fragile, sensitive to physical impact, and environmental conditions. Thus, a die or silicon chip is typically encapsulated in an epoxy or plastic to absorb and dissipate impact forces and to protect it from environmental conditions.

[0005] However, both underfilling and/or encapsulating a device or die increases the heat retained by the device. Maintaining a low operating temperature typically increases the reliability, performance, and life of an electronic device. Thus, the increase in temperature resulting from underfilling and/or encapsulation is an undesirable side effect of increasing the reliability by protecting an electronic device.

[0006] One type of semiconductor component is a memory component, which typically includes a memory die mounted on a substrate with or without encapsulation. Over the years, memory components have remained the same dimensional size while providing increased storage capacity. Generally, this has been accomplished by reducing the size of the individual storage elements on the memory component.

[0007] Memory components are often used in memory modules, where a number of these memory components are mounted on a single substrate. However, the number of memory components that may be placed on a substrate is limited by the size of the module. Modules must typically meet the functional and physical specifications established by industry standards or other limitations imposed by a particular application. Thus, the storage capacity of a memory module is often limited by physical size restrictions. In particular, the surface area available on the memory module limits the number of memory components that can be mounted thereon.

[0008] One technique to increase the storage capacity of memory modules has been to stack memory components on top of each other. This provides greater memory capacity for the same

surface area. However, the number of memory components that may be stacked is limited by the height restrictions on the memory module as well as the complexities of accessing stacked memory devices using existing bus interfaces designed for non-stacked architectures.

[0009] Thus, the prior art is still seeking an economical solution for increasing the stacking of memory devices while addressing the packaging requirements of memory modules.

SUMMARY OF THE INVENTION

[0010] One aspect of the invention provides a stackable semiconductor device architecture with each semiconductor device soldered onto a chip-scale ball grid array package. A second aspect of the invention provides a way to reduce thermal cracking of soldered points by using controlled thermal expansion substrates that substantially match the coefficient of expansion of semiconductor devices mounted thereon. A third aspect of the invention provides a semiconductor die mounting technique that improves heat dissipation by exposing all six surfaces of the semiconductor die mounted on a chip-scale package. A fourth aspect of the invention provides a chip-scale ball grid array package that permits mounting of components, such as capacitors and resistors, thereon. A fifth aspect the invention provides a staggered routing scheme that enables the use of the same trace routing at every level of the stacked architecture.

[0011] One implementation of the invention provides a memory module with stacked chip-scale ball grid array packages that increase the memory capacity of the module, while conforming to module dimensional requirements.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] Figure 1 illustrates a side view of a chip-scale ball grid array package with a semiconductor die mounted on a substrate according to one embodiment of the invention.

[0013] Figure 2 illustrates a surface view of a chip-scale ball grid array package with a semiconductor die mounted on a substrate according to one embodiment of the invention.

[0014] Figure 3 illustrates a side-view of a stacked pair of chip-scale ball grid array packages using memory dice according to one embodiment of the invention.

[0015] Figure 4 illustrates a surface view of a chip-scale ball grid array package with a semiconductor die mounted on a substrate according to a second embodiment of the invention.

[0016] Figure 5 illustrates a side-view of a stacked pair of chip-scale ball grid array packages using memory dice according to a second embodiment of the invention.

[0017] Figure 6 illustrates a memory module with stacked memory components on both surfaces of the memory module according to one embodiment of the invention.

[0018] Figure 7 illustrates a routing scheme for stackable packages according to one embodiment of the invention.

[0019] Figure 8 illustrates stacked memory components according to one implementation of the invention.

DETAILED DESCRIPTION

[0020] In the following description numerous specific details are set forth in order to provide a thorough understanding of the invention. However, one skilled in the art would recognize that the invention may be practiced without these specific details. In other instances, well known methods, procedures, and/or components have not been described in detail so as not to unnecessarily obscure aspects of the invention.

[0021] In the following description, certain terminology is used to describe certain features of one or more embodiments of the invention. For instance, "die" refers to a semiconductor device, such as a silicon memory device, that is not packaged or covered in a protective plastic, epoxy, or other material. The term "underside coupling members" is used to refer to such relatively rigid electrical coupling members as conductive bumps, conductive balls (e.g., solder or gold balls), and conductive rods.

[0022] One aspect of the invention provides a stackable semiconductor device architecture with each semiconductor device soldered onto a chip-scale package (CSP) having ball grid array connections. A second aspect of the invention provides a way to reduce thermal cracking of soldered points by using controlled thermal expansion substrates that substantially match the coefficient of expansion of semiconductor die mounted thereon. A third aspect of the invention provides a die mounting technique that improves heat dissipation by exposing all six surfaces of a die mounted in a chip-scale package. A fourth aspect of the invention provides a chip-scale package that permits mounting of signal conditioning (filtering) components, such as capacitors and resistors, thereon. A fifth aspect the invention provides a staggered routing scheme that enables the use of the same trace routing for every level of the stacked architecture.

[0023] One implementation of the invention provides a memory module with stacked memory components that increases the memory capacity of the module, while conforming to module dimensional requirements.

[0024] Figure 1 illustrates a side view of a chip-scale ball grid array package 100 including a semiconductor die 102 mounted on a substrate 104 according to one embodiment of the invention. The chip-scale ball grid array package 100 includes a plurality of solder balls 108 on a first surface of the substrate 104. One or more of the solder balls 108 may be electrically coupled to the semiconductor die 102 and/or contact pads 110.

[0025] One aspect of an implementation of the invention provides that the solder balls 108 have a higher profile than the semiconductor die 102 to mechanically protect the semiconductor die 102, from direct impact, etc., without the disadvantages of encapsulation. The solder balls 108 should be sufficiently large so that they rise above the semiconductor die 102 on the first surface of the chip-scale substrate 104. Mounting the semiconductor die 102 on the same side as solder balls 108 creates a flip chip assembly, where the semiconductor package can then be coupled to other substrates via the solder balls. Additionally, the higher profile solder balls permit the package 100 to be mounted on other substrates without interference from the semiconductor die 102.

[0026] Another aspect of the invention provides a way to reduce thermal cracking in the contacts between a semiconductor die and the substrate on which it is mounted. Conventional chip mounting techniques typically utilize wire bonds to electrically couple the semiconductor die to a contact point on the substrate. However, such mounting techniques typically require that the space between the semiconductor die and the mounting surface be underfilled or injected with a securing material, such as an epoxy, to fasten the semiconductor die to the substrate.

However, as previously discussed, this is undesirable because it tends to hinder heat dissipation from the semiconductor device.

[0027] As illustrated in Figure 1, a semiconductor die 102 is mounted on a substrate 104 using such rigid underside coupling members 106 as gold or solder balls, conductive polymer contacts, and/or under bump metallization. To prevent the coupling members 106 from cracking as a result of thermal cycling, one feature of the invention provides that the substrate 104 have a coefficient of thermal expansion matched to the coefficient of thermal expansion of the semiconductor die 102. That way, both the semiconductor die 102 and substrate 104 expand and contract at similar rates, thus reducing the stress on the coupling members 106. To achieve this, one aspect of the invention employs controlled thermal expansion (CTE) substrates that substantially match the coefficient of expansion of semiconductor devices mounted thereon.

[0028] Typical mounting substrates have a coefficient of expansion of approximately sixteen (16) to eighteen (18) parts per million per degree Celsius (ppm/°C) while silicon semiconductors have a nominal coefficient of expansion of approximately three (3) ppm/°C. Through temperature cycling, such disparity in coefficients of expansion often leads to cracks in electrical interconnects between such mounting substrates and silicon semiconductor devices. Employing a Coffin-Manson-type analysis, for instance, an adequate range of "matching" coefficients of expansion may be determined. An acceptable match will vary depending on the ranges of temperatures over which the materials will be cycled, the number of temperature cycles expected in the life of a device, the dimensions of the substrate, the material employed to couple a first substrate to a second substrate, etc. The closer the two materials (e.g., silicon semiconductor and mounting substrate) are to an exact match, the better the situation. Application-specific thermal-mechanical reliability may be readily tailored to meet the

requirements of the specific operating environment (e.g., solder and/or gold ball interconnections and/or conductive polymer materials).

[0029] In one implementation of the invention, a chip-scale ball grid array package includes a controlled CTE mounting substrate, having a coefficient of expansion of between six (6) and nine (9) ppm/°C, and a silicon semiconductor device having a nominal coefficient of expansion of approximately three (3) ppm/°C. The coefficient of expansion of the mounting substrate may be adjusted by varying its composition.

[0030] According to a Coffin-Manson-type analysis of a CTE material known as Thermount 85 NT, a coefficient of expansion of approximately eight (8) ppm/°C is acceptable for memory modules using silicon semiconductor devices having a coefficient of expansion of approximately three (3) ppm/°C. Such match is acceptable for dual in-line memory module (DIMM) applications, for instance.

[0031] Generally, the invention employs substrate materials made from a tailored CTE to assure the reliability for a given device in an operating environment. Some of the types of materials that may be used include:

- Arlon (Thermount) 85NT, with a coefficient of thermal expansion ranging from 6.0 to 9.0 ppm/°C, depending on the resin contents;
- Arlon (Thermount) 55 NT, with a coefficient of thermal expansion ranging from 7.0 to 10.0 ppm/°C, depending on the resin contents;
- CMC (copper-molybdenum-copper cores), with a baseline coefficient of thermal expansion greater than or equal to 5.5 ppm/°C, where varying the CTE by varying the copper thickness varies this coefficient (e.g., (copper-molybdenum-copper) 5/90/5 = 5.58 ppm/°C, 13/74/13 = 5.8 ppm/°C);

- CIC (copper-Invar-copper) cores, with a baseline coefficient of thermal expansion greater than or equal to 5.1 ppm/°C, where varying the CTE by varying the copper thickness varies this coefficient (e.g., (copper-Invar-copper) 20/60/20 = 5.2 ppm/°C);
- Ceramic cores having a coefficient of thermal expansion in the range of 4.5-6.5 ppm/°C (e.g., AlN at 4.5 and Alumina at 6.5).

[0032] Another aspect of the invention provides improved heat dissipation from the semiconductor die 102 by leaving all six surfaces of the die exposed. Unlike the prior art that is typically underfilled or completely encapsulated, the present invention exposes all six sides of a semiconductor die, including a substantial portion of the underside of the semiconductor die. That is, by mounting the semiconductor die 102 using solder balls, under bump metallization, and/or other similar electrically coupling members 106, a gap is created between the chip-scale substrate 104 and the underside of the semiconductor die 102. Because all of the surfaces of the die 102, including the underside surface, are now exposed to airflow, the semiconductor die 102 has improved heat dissipation. Note that the fact that the solder balls 108 have a higher profile than the mounted semiconductor die 102 means that the upper surface of the semiconductor die 102 is also exposed to airflow.

[0033] Figure 2 illustrates a surface view of a chip-scale ball grid array package utilizing a semiconductor die mounted on a substrate according to one embodiment of the invention. The chip-scale ball grid array package may include a plurality of solder balls 108 along a surface of the substrate 104. In one implementation, the plurality of solder balls 108 may be arranged in rows and columns, as shown in Fig. 2, while in other implementations the plurality of solder balls 108 may be arranged in other configurations.

[0034] Figure 3 illustrates a stacked chip-scale ball grid array package using memory dice according to one embodiment of the invention. A plurality of the chip-scale ball grid array packages 302 and 304 may be vertically stacked with a first surface of a first package 302 being coupled to an opposing second surface of a second package 304 and so on for each successive layer in the stack. In particular, the solder balls 306 on the second surface of the second package 304 are electrically coupled to corresponding pads (e.g., 110 in Fig. 1) on the first surface of the first package 302. Because the solder balls 306 have a larger vertical profile or height than the semiconductor die 305, this permits stacking the first package 302 over the second package 304. In this manner, a plurality of packages may be stacked to increase the density of semiconductor devices that may be mounted on a given area. For example, when the semiconductor die 301 and 305 are memory die, stacking a plurality of memory devices increases the capacity of a memory module in comparison to single-layer chip architectures.

[0035] One aspect of the invention provides a chip-scale ball grid array package that permits mounting of components, such as capacitors and resistors, thereon. By mounting the semiconductor die 305 on the substrate 304 using connects 303, surface space is freed on the substrate above semiconductor die 305. In one implementation, the surface space above the semiconductor die always includes pads 308 on which signal conditioning components may be mounted. This surface area may have one or more pads 308 for connecting signal filtering components thereon. This permits mounting on-chip electrical components 310, such as capacitors and resistors, which may be used for signal conditioning to and/or from the semiconductor die 301. Being able to mount components 310 on the package substrate itself (e.g., chip-scale substrate) is an advantage over the prior art that was limited to mounting said

components only external to the package. In one implementation of the invention, such components 310 are signal conditioning capacitors and pull-up/pull-down resistors.

[0036] According to other implementations, other ball grid array configurations may be employed without deviating from the invention.

[0037] Figure 4 illustrates a surface view of a chip-scale ball grid array package with a semiconductor die mounted on a substrate according to a second embodiment of the invention. This configuration includes a substrate 402 with a plurality of solder ball connects 404 and 406 arranged as two columns on either side of a semiconductor die 408 mounted on the substrate 402. The semiconductor die 408 may be mounted on the substrate 402 by rigid underside coupling members 410 and 412, also arranged in two columns along the length of the semiconductor die 408. The underside coupling members 410 and 412 are electrically coupled to a plurality of the solder balls 404 and 406. The solder balls 404 and 406 are electrically coupled to pads on the opposite surface of the substrate 402.

[0038] In one implementation of the invention, the outboard columns of balls/electrical interconnections (e.g., 404 and 406) is two wide per side times whatever length is required to effectively address the maximum number of chip-scale packages to be stacked. For example, the basic I/O count required for a typical two hundred and fifty-six (256) Megabyte (Mb) synchronous dynamic (SD) random access memory (RAM) silicon device is forty-eight (48), two of which are for unique "addressing". Hence, in a one high implementation, only forty-eight (48) I/O ball/electrical interconnections 404 and 406 would be required. Since each additional device requires unique clock enable and chip select addressing interconnects, a two high stack would require that all chip-scale packages in that stack have fifty (50) interconnects 404 and 406.

A four-high stack would require fifty-four (54) ball/electrical interconnections 404 and 406 and so forth.

[0039] Figure 5 illustrates a side-view of stacked chip-scale ball grid array packages using memory dice according to a second embodiment of the invention. A plurality of chip-scale packages 502 and 504, like that illustrated in Fig. 4, are stacked as shown and in a similar manner to the stacked package described in Fig. 3. Each substrate 502 and 504 includes two columns of solder balls 506 along the sides of semiconductor devices 510 and 512 mounted thereon. A plurality of underside electrical interconnects 508 couples the semiconductor devices 510 and 512 to their respective substrates 502 and 504. Each substrate also includes a plurality of pads 514 on the surface opposite the semiconductor device that may be used to couple signal filtering components 516.

[0040] Figure 6 illustrates a memory module 600 with a plurality of stacked memory components 602-612 on both surfaces of the memory module 614 according to one embodiment of the invention. A plurality of stacked chip-scale ball grid array packages 602-612 may be mounted on one or more surfaces of a substrate 614.

[0041] In one implementation of the invention, the semiconductor devices (e.g., 102) may be random access memory devices mounted on stacked chip-scale packages (e.g., 602-612). The stacked packages (e.g., 602-612) are then mounted on either or both sides of a substrate to form a memory module 600, such as a single inline memory module (SIMM) or dual inline memory module (DIMM). The dimensional requirements of the memory module 600 may limit the number of packages (e.g., 100) that may be stacked.

[0042] In yet other implementations, the stacked packages may be mounted directly on a computer motherboard or other type of module.

[0043] Figure 7 illustrates a routing scheme for a package stack 700 according to one embodiment of the invention. One difficulty in implementing the stackable package architecture is providing a simple way to access each semiconductor device (e.g., memory device) independently. Prior art stackable architectures have typically required customizing the chip-scale package substrate at each level of a stack to enable access to each chip independently. That is, for the same pin on each a chip in a package stack, a different electrical contact must be used at the interface with the module substrate to access each chip independently. This typically necessitates customizing chip-scale package substrates at each level of the stacked packages, with jumpers or trace punching for instance, to provide the appropriate routing from the interface with the module board to the chip. However, manufacturing and assembling such stacked semiconductor chip-scale packages is costly and cumbersome since customized chip-scale package substrates are typically necessary for each level of the stacked package.

[0044] According to one embodiment of the invention, identical chip-scale package substrates 701a-d are employed at each level of the stacked package 700 with the semiconductor dice 703a-d mounted in the same manner on each chip-scale substrate 701a-d. Using identical substrates 701a-d for every level of the stacked package 700 and mounting the dice 703a-d in an identical manner avoids the costs and difficulties of having to customize each level. Thus, a plurality of chip-scale ball grid array packages 701a-d may be manufactured and later assembled into a stack without regard as to any particular order.

[0045] The use of identical chip-scale packages at every level of a stack package 700 is made possible by a novel routing scheme that permits accessing each semiconductor die 703a-d independently and without customization of the substrates 701a-d. The novel routing scheme provides cascading connections through all levels of the stack package 700 to electrically couple

each semiconductor die 703a-d to a primary access point (e.g., the solder balls on substrate 701a).

[0046] According to implementation of the novel routing scheme, each substrate 701 includes a plurality of solder balls 705 mounted on a first surface of the substrate 701 and a plurality of corresponding pads 707 on a second opposite surface of the substrate 701. Interconnects 709 serve to electrically couple the solder balls 705 to the pads 707. A cascading scheme, as illustrated in Fig. 7, permits independently accessing each semiconductor die from a primary access point (e.g., the solder balls on substrate 701a). The cascading scheme electrically couples the solder balls closest to the semiconductor die to a contact on the semiconductor die. For example, solder ball 711 is electrically coupled to a contact 702 on semiconductor die 703a. This implementation of the routing scheme provides for solder balls, on a first surface of a chip-scale package substrate, to be electrically coupled to pads, on a second surface of the chip-scale package substrate, that are closer to the semiconductor die. For example, solder ball 713 is electrically coupled to pad 715. Similarly, solder balls 717 and 719 are coupled to pads that are closer to the semiconductor die 703a. This routing scheme is implemented at every chip-scale package substrate 701a-d in the stack package 700. When the chip-scale package substrates are stacked on top of each other as shown, the each pad (e.g., 707) on the second surface of a first substrate (e.g., 701a) is electrically coupled to a corresponding solder ball (e.g., 720) on a first surface of a second substrate (e.g., 701b). Thus, a cascading routing scheme results that electrically couples solder balls (e.g., 713) in one substrate (e.g., 701a) to solder balls (e.g., 722) in a second substrate (e.g., 701b) that are closer to the interface point of a semiconductor die. Consequently, solder ball 713 is electrically coupled to connect 706, solder ball 717 is electrically coupled to connect 710, and solder ball 719 is electrically coupled to connect 714. A

similar scheme may be implemented on the other side of the semiconductor die such that die connects 704, 708, 712, and 716 are electrically coupled to solder balls on substrate 701a. The stacked package of semiconductor devices may then be coupled to another substrate or interface via the solder balls on the first surface of substrate 701a.

[0047] Note that the solder balls may be cascaded in different ways and utilizing different solder ball layouts without deviating from the invention. For example, a cascading routing scheme may be implemented using chip-scale packages similar to that illustrated in Figures 4 and 5. The cascading scheme may be implemented in two solder ball columns (e.g., 404) by zigzagging back and forth between solder balls across multiple package layers to effectively implement the routing scheme illustrated in Fig. 7. That is, the electrical pad on a first surface of a substrate is electrically coupled to a solder ball on a second surface of the substrate and substantially diagonal relative to the alignment of other pads and solder balls. In other implementations, the routing scheme shown in Fig. 7 may also be implemented by interconnecting solder balls along a single column (e.g., one of the outer columns in semiconductor package 400) across multiple chip-scale package layers. In short, the staggered routing scheme illustrated in Figs. 7 and 8 may be implemented in various ways, including single column, single row, and/or diagonal/zigzag interconnections across various layers of a stack.

[0048] In yet other implementations, only part of the solder balls are interconnected using a cascading scheme while part of the remaining solder balls are connected in a non-cascading manner across the stacked package. That is, some of the solder balls in the same position across all layers of a stack may be commonly connected.

[0049] Figure 8 illustrates a stacked memory package 800 according to one implementation of the invention. Memory dice 802a-d are mounted on chip-scale package substrates 804a-d. In

one implementation, the memory components 802a-d share common data lines or a bus that serves to write and/or read data (e.g., bits) to and from the memory devices 802a-d. In order to correctly write to and/or read from a particular memory device, the cascading scheme described in Fig. 7 is implemented for individual chip selection and clock enabling. The individual packaged memory dice 802a-d may be enabled and/or disabled providing an appropriate signal at solder balls 806-813. For example, memory device 802a may be accessed by enabling Clock A at solder ball 806 and chip Select A at solder ball 807. Similarly, memory device 802b may be accessed by enabling Clock B at solder ball 808 and chip Select B at solder ball 809, memory device 802c may be accessed by enabling Clock C at solder ball 810 and chip Select C at solder ball 811, and memory device 802d may be accessed by enabling Clock D at solder ball 812 and chip Select D at solder ball 813.

[0050] Oftentimes, the size of a stacked memory or semiconductor package is limited by the space available on which to mount it in a particular implementation. Thus, one implementation of the invention employs tightly spaced components to maximize semiconductor or memory density of a stack package. For instance, in one implementation of the invention, the substrate (e.g., 104 in Fig. 2 or 402 in Fig. 4) is approximately twelve (12) millimeters (mm) long by ten (10) mm wide by 0.20 mm thick. Additionally, the solder balls (e.g., 308 in Fig. 3, or 506 in Fig. 5) may be approximately 0.50 mm in diameter, while the height or thickness of a particular chip-scale package (e.g., 100 in Fig. 1) is approximately 0.50 mm. Similarly, the thickness or height of two stacked chip-scale packages (e.g., as those illustrated in Figs. 3 and 5) is approximately 1.00 mm and so forth.

[0051] While certain exemplary embodiments have been described and shown in the accompanying drawings, it is to be understood that such embodiments are merely illustrative of

and not restrictive on the broad invention, and that this invention not be limited to the specific constructions and arrangements shown and described, since various other modifications are possible. Those skilled, in the art will appreciate that various adaptations and modifications of the just described preferred embodiment can be configured without departing from the scope and spirit of the invention. For example, while a semiconductor die has been used to illustrate the invention, any other electronic device or component may be used instead, with one or more aspects of the invention, without deviating from the invention. Therefore, it is to be understood that, within the scope of the appended claims, the invention may be practiced other than as specifically described herein.